



CHALMERS
UNIVERSITY OF TECHNOLOGY



MATLAB-Based Implementation of a Spatial Covariance Matching Spread Rendering Method

Master's Thesis in Civil and Environmental Engineering

Zhengxiang Zhou

DEPARTMENT OF Civil and Environmental Engineering

CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden, 2026
www.chalmers.se

MASTER'S THESIS 2026

MATLAB-Based Implementation of a Spatial Covariance Matching Spread Rendering Method

Zhengxiang Zhou



DEPARTMENT OF Civil and Environmental Engineering
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden, 2026

An Implementation and Analysis of Covariance-Domain Source Spread Rendering
in MATLAB

Zhengxiang Zhou

© Zhengxiang Zhou, 2026.

Supervisor: Jens Ahrens, Department of Civil and Environmental Engineering
Examiner: Jens Ahrens, Department of Civil and Environmental Engineering

Master's Thesis 2026
Department of Civil and Environmental Engineering
Chalmers University of Technology
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Typeset in L^AT_EX
Gothenburg, Sweden, 2026

An Implementation and Analysis of Covariance-Domain Source Spread Rendering in MATLAB

Zhengxiang Zhou

Department of Civil and Environmental Engineering

Chalmers University of Technology

Abstract

A MATLAB implementation of the spatial covariance matching method proposed by McCormack et al. is developed for the case of binaural reproduction and implemented in an afSTFT framework using HRTFs, alongside baseline rendering (BL) and eigenvalue decomposition (EVD) and optimal mixing (OM). In the OM method, both prototype and residual mixing matrices are generated by the covariance-domain optimization process as introduced in the original method. The evaluation of performance is done in terms of colouration, interaural level difference (ILD), interaural coherence (IC), RMSE analysis over 836 source directions at different source spreads, and signal distortion error (SDE). The results indicate that EVD improves coherence matching and spatial diffuseness and OM yields a compromise between spatial widening and signal preservation. The global performance is consistent overall with the reference implementation, but differences exist between the implementations for certain metrics and signal types in terms of quantity.

Keywords: spatial covariance matching, source spread rendering, eigenvalue decomposition, optimal mixing, afSTFT, HRTF, binaural rendering, interaural coherence, interaural level difference, signal distortion error

Acknowledgements

This master thesis was developed under the supervision and support of Jens Ahrens. Zhengxiang Zhou acknowledges the support and encouragement of his wife, Dingqianxi Zhu, throughout the project.

Zhengxiang Zhou, Gothenburg, May 2026

List of Acronyms

Below is the list of acronyms used throughout this thesis, listed in alphabetical order:

afSTFT	Adaptive/Analysis Filterbank Short-Time Fourier Transform
EVD	Eigenvalue Decomposition
HRTF	Head-Related Transfer Function
IC	Interaural Coherence
ILD	Interaural Level Difference
OM	Optimal Mixing
SCM	Spatial Covariance Matrix
SDE	Spatial Deviation Error
STFT	Short-Time Fourier Transform

Nomenclature

Below is the nomenclature of indices, sets, parameters, and variables used throughout this thesis.

Indices

q	Ear channel index (left/right)
f	Frequency-band index
t	Time-frame index
s	Source index
d	Spatial direction index
i	Hybrid-band index
n	Time-sample index in time-domain signals

Sets

$\mathcal{S}$	Set of sources
$\mathcal{Q}$	Set of ear channels
$\mathcal{F}$	Set of frequency bands
$\mathcal{T}$	Set of time frames
$\mathcal{A}$	Spatial spreading region on the unit sphere

Parameters

$x_q[n]$	Time-domain input signal at ear channel q and sample n
$w_a[n]$	Analysis window used in afSTFT analysis
$w_s[n]$	Synthesis window used in afSTFT reconstruction

$\mathbf{C}_y(f)$	Target spatial covariance matrix of the rendered binaural signals at frequency band f
$\mathbf{D}(f)$	Spread diffuse coherence matrix (SDCM) at frequency band f
$\mathbf{C}_{\text{proto}}(f)$	Prototype covariance matrix used in OM rendering
$\mathbf{C}_{\text{OM}}(f)$	Spatial covariance matrix of the OM-rendered signal
$\mathbf{C}_{\text{dec}}(f)$	Covariance matrix of the decorrelated residual signal
$\mathbf{C}_r(f)$	Residual covariance matrix used for covariance compensation in OM rendering
$\mathbf{h}(f, \mathbf{u})$	Directional response vector for direction $\mathbf{u}$ at frequency band f
$\mathbf{u}$	Direction-of-arrival unit vector
w_k	Integration weight for directional sampling point k
$P_s(f)$	Source power at frequency band f
$\mathbf{M}(f)$	Covariance-domain mixing matrix used in EVD and OM rendering
$\mathbf{M}_r(f)$	Residual mixing matrix used in OM rendering
$\mathbf{\Lambda}(f)$	Diagonal eigenvalue matrix obtained from EVD of the diffuse coherence matrix
$\mathbf{E}(f)$	Eigenvector matrix obtained from EVD of the diffuse coherence matrix
$\mathcal{D}[\cdot]$	Decorrelation operator applied to the prototype signal
K	Total number of frequency bands used in RMSE evaluation
x_k	Evaluated perceptual metric at frequency band k
$x_{\text{ref},k}$	Reference perceptual metric at frequency band k
$c_{d,ll}(f)$	Auto-spectral power of the left binaural channel
$c_{d,rr}(f)$	Auto-spectral power of the right binaural channel
$c_{d,lr}(f)$	Cross-spectral term between left and right binaural channels
$c_{d,rl}(f)$	Cross-spectral term between right and left binaural channels
c_1, c_2, c_3	Hybrid-band weighting coefficients
frameSize	Frame length in samples
hopSize	Hop size in samples
nBands	Number of frequency bands including hybrid bands
nSources	Number of sources
$\text{Colouration}(f)$	Binaural colouration metric at frequency band f
$\text{ILD}(f)$	Interaural level difference at frequency band f
$\text{IC}(f)$	Interaural coherence at frequency band f
RMSE	Root-mean-square error between rendered and reference perceptual metrics

SDE	Signal distortion error between rendered and reference binaural signals
Q	Number of output channels

Variables

$\mathbf{d}(f, t)$	General binaural frequency-domain signal vector at frequency band f and frame t
$\mathbf{d}_{\text{dec}}(f, t)$	Decorrelated prototype signal generated using the decorrelation operator
$\mathbf{d}_{\text{proto}}(f, t)$	Prototype binaural signal vector used in OM rendering
$\mathbf{d}_{\text{OM}}(f, t)$	Output binaural signal vector generated by OM rendering
$\mathbf{d}_{\text{scm}}(f, t)$	Rendered frequency-domain signal vector after covariance-domain mixing
$X_{f,q}(t)$	Frequency-domain coefficient for channel q at frequency band f and frame t
$Y_{i,q}(t)$	Hybrid-band coefficient generated from low-frequency frequency bins
$X_{i,q}^{\text{bin}}(t)$	Reconstructed frequency-bin coefficient after hybrid-band recombination
$y_q[n]$	Reconstructed time-domain signal at ear channel q and sample n
$\mathbf{d}_{\text{ref}}(n)$	Reference binaural signal in the time domain

Contents

List of Acronyms	ix
Nomenclature	xi
1 Introduction	1
1.1 Motivation and Significance of Spatial Audio	1
1.2 Challenges in Implementing Covariance-Domain Rendering	1
1.3 Related Works	2
1.4 Objectives	2
1.5 Contributions	3
1.6 Organization of the Thesis	3
2 Theory	5
2.1 Spatial Audio and Perceptual Metrics	5
2.2 Short-Time Fourier Transform and Analysis-Synthesis	6
2.3 Covariance-Domain Source Spread Rendering	7
2.3.1 Spatial Covariance Matrices	7
2.3.2 Eigenvalue Decomposition Rendering	9
2.3.3 Optimal Mixing Rendering	10
2.4 HRTF-Based Directional Modeling	12
2.5 Decorrelation Mechanisms	13
3 MATLAB Implementation and System Framework	15
3.1 Overall Rendering Pipeline	15
3.2 afSTFT Processing Framework	16
3.3 HRTF Processing and Directional Grid Construction	18
3.4 EVD Rendering Implementation	19
3.5 Optimal Mixing Rendering	20
4 Results and Discussion	23
4.1 Evaluation Methodology	23
4.1.1 Evaluation Signals and Rendering Conditions	23
4.1.2 Binaural Covariance Representation	23
4.1.3 Perceptual Metrics	23
4.1.4 RMSE-based Metric Evaluation	24
4.1.5 Signal Distortion Error	24
4.2 Objective Evaluation Results	25

4.2.1	Frequency-Dependent Binaural Metric Analysis	25
4.2.2	Spread-Dependent Metric Analysis	26
4.2.3	Spatial Deviation Error Analysis	28
4.3	Comparison with the Original Results	30
4.3.1	Frequency-Dependent Metric Comparison	30
4.3.2	Global RMSE Comparison Across Source Spreads	31
4.3.3	Spatial Deviation Error Comparison	33
5	Conclusion and Future Work	35
	Bibliography	37
A	Core MATLAB Implementation Snippets	I
A.1	EVD Covariance-Domain Mixing	I
A.2	OM Covariance-Domain Mixing	I
B	Rendering and Evaluation Parameters	III

1

Introduction

1.1 Motivation and Significance of Spatial Audio

The ability to accurately localize a sound source, separate overlapping sounds and keep a coherent spatial scene [1], [2] is essential for human auditory perception. In immersive environments like virtual reality (VR), augmented reality (AR), and 3D multimedia, spatial fidelity plays a crucial role in the overall perceptual realism and the sense of presence for the user, and it is particularly important to ensure accurate representation of these cues in such contexts [2].

The traditional way to render a spatial image, such as amplitude panning, vector based amplitude panning [3] and ambisonics, are aimed mainly at positioning discrete sound sources. Although adequate for the localization task, these methods provide a much less precise control on perceived spatial extent, and they might affect timbral fidelity and interaural coherence [4]. Simple signal decorrelation techniques can help increase perceived spread, but may also involve artifacts, especially in multi-source scenes.

The extension of the control over the spatial properties is achieved in the covariance domain source spread rendering where the spatial covariance matrix of the sound field is directly manipulated [5]. Such an approach allows the ability to adjust the spatial spread and preserve the directional cues and natural timbre, which offers an effective basis to assess and apply EVD and OM strategies. Such a formulation provides the conceptual groundwork for studying practical problems and design issues in robust spatial audio.

1.2 Challenges in Implementing Covariance-Domain Rendering

While covariance-domain rendering allows for a systematic control of the source spread, there are several problems in its implementation. For EVD rendering, decorrelated signal components are combined to generate a desired covariance between them. This method may lead to good spatial diffuseness but can also cause decorrelation in the spectral and binaural qualities of the reproduced signal.

A more constrained alternative is provided by Optimal Mixing (OM), where a prototype component is used together with a decorrelated residual component. The mixing matrices in this case need to be calculated so that the sum of the prototype and residual branches brings up the target covariance close to the value. These are

some of the factors to consider when implementing the model: accurate estimation of the covariance, numerical stability and handling residual covariance.

Other problems occur with frame-based processing and binaural rendering based on a HRTF. The frequency-dependent mixing matrices have to change smoothly from one processing frame to the next, and the directional HRTF response has to be consistently represented in the time-frequency domain. Validation is also required to determine whether the MATLAB implementation reproduces the spatial and spectral behaviour reported by McCormack et al.[5] In case of evaluation the following parameters are applied: Colouration, interaural level difference (ILD), interaural coherence (IC), signal distortion measures.

1.3 Related Works

There has been prior work which has investigated different approaches to spatial audio reproduction. The early techniques include vector-based amplitude panning and virtual source positioning [3], but they have limited control over perceptual width and can cause some loss of timbre fidelity. It is possible to enhance the perceived spatial spread by simple decorrelation, but this can also impact localization accuracy and overall sound quality [6].

The covariance domain methods solve these problems by directly transforming the spatial covariance matrix, instead of the audio signals. McCormack et al. [5] were able to formalize a framework that combines Eigenvalue Decomposition (EVD) with Optimal Mixing (OM) with target covariance matching. Here, EVD decorrelates all the sources to the fullest extent, whilst the OM structure preserves the prototype structure and introduces an additional decorrelated component to achieve the balance of timbre and spread. This framework provides a theoretical basis for the exact control of spatial properties and perceptual dispersion.

In binaural rendering, Head-Related Transfer Functions (HRTFs) play an essential role in the direction-dependent filtering that mimics the acoustical properties of a listener’s head, torso, and pinnae. By reproducing binaural spatial cues through HRTF-based processing, spatial audio systems can improve localization performance, externalization, and perceptual realism [7]. Extensions of covariance-domain methods have been developed for arbitrary playback setups, including the management of residual energies, the frame-wise interpolation of mixing matrices and the decorrelation versus audibility compromise.

1.4 Objectives

The main goal of this thesis is to implement and evaluate a MATLAB-based reproduction of the covariance-domain source spread rendering framework proposed by McCormack et al. [5]. The objectives are as follows:

- **Reproduce EVD and OM rendering strategies:** Implement frame-wise processing and afSTFT analysis-synthesis to match the original methods.
- **Integrate HRTF-based directional modeling:** Ensure binaural outputs accurately represent spatial cues for localization and spread perception.

- **Evaluate spatial and perceptual performance:** Use objective metrics (interaural coherence, interaural level differences, coloration) to compare MATLAB reproduction with original results.
- **Compare the MATLAB implementation with the reference method:** identify similarities and remaining quantitative differences in the rendering behaviour.
- **Analyse implementation trade-offs:** examine the relationship between covariance matching, spatial diffuseness, binaural cue preservation, and signal distortion.

1.5 Contributions

This thesis provides several contributions:

1. A modular MATLAB implementation of the covariance-domain source spread rendering, which allows it to be replicated and used for future research.
2. Systematic comparison of EVD and OM strategies emphasizing the differences in spatial accuracy, timbre preservation and perceptual quality.
3. Practical implementation considerations for covariance-domain rendering, including decorrelation processing, residual covariance estimation, positive semi-definite projection, and covariance-domain mixing, are documented and analysed.

1.6 Organization of the Thesis

The rest of the thesis is organized as follows. Chapter 2 presents the theoretical background of covariance-domain spatial rendering, including spatial covariance matrices, EVD and OM rendering, afSTFT processing, and HRTF-based directional modeling. Chapter 3 describes the overall system framework and MATLAB implementation of the rendering pipeline. Chapter 4 presents the evaluation methods, rendering results, and discussion of the implementation differences and perceptual trade-offs between the present framework and the original method. Finally, Chapter 5 concludes the thesis and discusses limitations and possible future work.

2

Theory

2.1 Spatial Audio and Perceptual Metrics

Precise determination of interaural cues is vital to human auditory perception for locating the source of a sound and for understanding an acoustic scene. The similarity of the signals measured at the left and right ear is called interaural coherence (IC), which quantifies spatial spread and signal decorrelation. Interaural level difference (ILD) is the difference in sound level between ears and is a good approximation of directional perception in the horizontal plane. The colouration is an indicator of the spectral fidelity of the reproduced sound compared to the original sound, which represents the timbre preservation characteristics of spatial processing [8]. The perceptual aspects are key to the assessment of binaural rendering systems and the design of spatial audio algorithms.

Head-Related Transfer Functions (HRTFs) are signal models that describe the direction-dependent filtering effects of the listener’s outer ears, head and torso. Each HRTF conveys a frequency-dependent transfer function from a specific source direction to each ear, which can be used to compute the other HRTF cues required to accurately perceive azimuth and elevation [5]. HRTFs are typically recorded or simulated on high-resolution space sampling grid and interpolated for arbitrary directions of the sound source. The incorporation of HRTFs for spatial audio rendering allows development of a accurate binaural perception and maintaining perceptual integrity for multiple concurrent sources.

Spatial covariance represents the second order statistical correlation between channels or ears and is a formal representation of the spatial structure of multiple sources. It is possible to modify the covariance matrix, thereby controlling the spatial dimensions of the sources and the directionality of the signals rendered without changing the individual source signals. This property can be used to improve perceptual spread without compromising localization accuracy and timbral fidelity in covariance-domain approaches [5], [9].

In immersive applications like VR, AR, and 3D multimedia, spatial reproduction is crucial to ensure perceptual realism and sense of presence [2], [10]. Together, the use of interaural cues, HRTFs and covariance-domain control helps listeners determine direction and spatial width of multiple sources and ensures natural perception of audio across various virtual environments.

2.2 Short-Time Fourier Transform and Analysis-Synthesis

Spread rendering is done in the covariance domain, operating in the time-frequency domain and enabling processing of the spatial spread characteristics of the sources frame-by-frame. It is implemented based on the framework of the alias-free short-time Fourier transform (afSTFT) that features a windowed overlap-add (WOLA) structure for analysis and synthesis. The afSTFT compares with a conventional STFT and also adds hybrid low-frequency bands that provides better low-frequency resolution but retain the same reconstruction quality.

Let $x_q[n]$ be a time domain amplitude of an ear channel q input signal. When using the frame-based processing, all the samples received in a frame are added to the analysis buffer and the overlapping samples of the previous frame are saved so that there is temporal continuity. The signal after having been buffered is multiplied by an analysis window before the time-frequency transformation is applied:

$$x_q^{\text{win}}(n, t) = x_q^{\text{buf}}(n, t) \cdot w_a[n], \quad (2.1)$$

where $x_q^{\text{buf}}(n, t)$ represents the buffered frame for channel q at frame index t , and $w_a[n]$ denotes the analysis window.

The frequency-domain coefficients are obtained using the discrete Fourier transform:

$$X_{f,q}(t) = \text{FFT} \{ x_q^{\text{win}}(n, t) \}, \quad (2.2)$$

where f denotes the frequency-bin index. Only the positive-frequency bins are retained after transformation. The resulting coefficients are normalized using the sum of the analysis window to maintain amplitude consistency across frames.

To improve low-frequency resolution, hybrid processing is applied to the first four frequency bins. For each of these bins, two hybrid sub-bands are generated using the current frame and two previous frames:

$$Y_{2i-1,q}(t) = \frac{c_3 X_{i,q}(t) + c_2 X_{i,q}(t-1)}{g}, \quad (2.3)$$

$$Y_{2i,q}(t) = \frac{c_1 X_{i,q}(t-2)}{g}, \quad (2.4)$$

where $i = 1, \dots, 4$, c_1 , c_2 , and c_3 are the hybrid weighting coefficients, and $g = c_1 + c_2 + c_3$ is the normalisation factor. The remaining frequency bins are copied directly into the hybrid representation:

$$Y_{i+4,q}(t) = X_{i,q}(t), \quad i = 5, \dots, 129. \quad (2.5)$$

Thus, the first four frequency bins are expanded into eight hybrid bands, The remaining higher-frequency bins are copied directly into the hybrid representation without modification. This increases the total number of analysis bands from 129 to 133.

After frequency-domain processing, the inverse afSTFT reconstructs the time-domain signal. If hybrid processing is enabled, the additional low-frequency bands are first

merged back into the original frequency-bin representation. The low-frequency bins are reconstructed by recombining the corresponding hybrid bands:

$$X_{i,q}^{\text{bin}}(t) = Y_{2i-1,q}(t) + Y_{2i,q}(t), \quad i = 1, \dots, 4. \quad (2.6)$$

The remaining bins are shifted back directly. A conjugate-symmetric spectrum is then constructed before inverse Fourier transformation:

$$y_q^{\text{frame}}(n, t) = \text{IFFT} \left\{ X_q^{\text{bin}}(f, t) \right\}. \quad (2.7)$$

The reconstructed frame is multiplied by the synthesis window and accumulated in the synthesis buffer using overlap-add:

$$y_q[n] = \sum_t y_q^{\text{frame}}(n, t) \cdot w_s[n - tR], \quad (2.8)$$

where R denotes the hop size and $w_s[n]$ is the synthesis window. The synthesis buffer outputs one hop-length segment at each frame, producing a continuous time-domain signal. This WOLA-based analysis-synthesis structure maintains stable frame continuity and provides near-perfect reconstruction for subsequent covariance-domain processing.

2.3 Covariance-Domain Source Spread Rendering

2.3.1 Spatial Covariance Matrices

Consider a binaural rendering system with Q output channels. In general, the rendering system can be characterised by a directional response vector that depends on both frequency and direction of arrival. For binaural headphone rendering, these directional responses correspond to head-related transfer functions (HRTFs), while for loudspeaker-based rendering they may correspond to panning gains or array responses. The directional response vector for a direction $\mathbf{u}$ and frequency band f is written as

$$\mathbf{h}(f, \mathbf{u}) = \begin{bmatrix} h_1(f, \mathbf{u}) \\ h_2(f, \mathbf{u}) \\ \vdots \\ h_Q(f, \mathbf{u}) \end{bmatrix}, \quad (2.9)$$

where $\mathbf{u}$ denotes the direction of arrival on the unit sphere. These directional responses may be obtained analytically, modelled numerically, or measured directly from a dense spatial grid surrounding the listener. In the binaural case used in this implementation, the directional responses are derived from measured HRTF datasets.

The directional response vectors for multiple spatial directions can be combined into a directional response matrix

$$\mathbf{H}(f) = [\mathbf{h}(f, \mathbf{u}_1), \mathbf{h}(f, \mathbf{u}_2), \dots, \mathbf{h}(f, \mathbf{u}_K)], \quad (2.10)$$

where K denotes the total number of spatial grid directions.

A spatially spread sound source occupying the angular region $\mathcal{A}$ can be modelled as a superposition of mutually incoherent directional source components. The resulting binaural signal vector is expressed as

$$\mathbf{d}(f) = \int_{\mathbf{u} \in \mathcal{A}} \mathbf{h}(f, \mathbf{u}) s(f, \mathbf{u}) d\mathbf{u}, \quad (2.11)$$

where $s(f, \mathbf{u})$ denotes the source signal arriving from direction $\mathbf{u}$ and $d\mathbf{u}$ denotes the differential surface element on the unit sphere. The integration is performed over all directions within the spreading region $\mathcal{A}$.

The covariance-domain formulation assumes that the directional source components are mutually incoherent:

$$\mathbb{E}[s(f, \mathbf{u}_i) s^*(f, \mathbf{u}_j)] = \begin{cases} 0, & \mathbf{u}_i \neq \mathbf{u}_j, \\ P_s(f, \mathbf{u}_i), & \mathbf{u}_i = \mathbf{u}_j, \end{cases} \quad (2.12)$$

where $P_s(f, \mathbf{u})$ denotes the directional source power distribution and $\mathbb{E}[\cdot]$ represents the expectation operator. This assumption implies that directional components arriving from different directions are statistically uncorrelated.

Under this assumption, the spatial covariance matrix (SCM) of the rendered binaural signal becomes

$$\mathbf{C}_y(f) = \mathbb{E}[\mathbf{d}(f) \mathbf{d}^H(f)] = P_s(f) \mathbf{D}(f), \quad (2.13)$$

where $\mathbf{D}(f)$ denotes the spread diffuse coherence matrix (SDCM) associated with the rendering system. The SDCM describes the diffuse spatial coherence structure produced by the directional responses of the renderer.

The SDCM is defined as

$$\mathbf{D}(f) = \int_{\mathbf{u} \in \mathcal{A}} \mathbf{h}(f, \mathbf{u}) \mathbf{h}^H(f, \mathbf{u}) d\mathbf{u}. \quad (2.14)$$

For practical implementations, the continuous integral is approximated numerically using a discrete spatial grid:

$$\mathbf{D}(f) \approx \sum_{k=1}^K w_k \mathbf{h}(f, \mathbf{u}_k) \mathbf{h}^H(f, \mathbf{u}_k), \quad (2.15)$$

where w_k denotes the integration weight associated with direction $\mathbf{u}_k$. In the present implementation, the directional response vectors are obtained from measured HRTFs distributed across a discrete spatial grid. The weighted summation of directional covariance contributions forms the target diffuse coherence structure used for covariance-domain rendering.

The target covariance matrix $\mathbf{C}_y(f)$ defines the desired spatial spread characteristics of the rendered source. Subsequent rendering methods, including eigenvalue decomposition (EVD) and omnidirectional mixing (OM), construct mixing matrices such that the covariance of the output signals approximates the target covariance structure.

2.3.2 Eigenvalue Decomposition Rendering

The objective of covariance-domain rendering is to construct output signals whose spatial covariance matches a desired target covariance structure. In the present framework, the target spatial characteristics are represented by the spread diffuse coherence matrix (SDCM) $\mathbf{D}(f)$ derived from the directional rendering responses. Eigenvalue decomposition (EVD) rendering addresses this problem by transforming mutually decorrelated input components into output signals whose covariance follows the target diffuse coherence structure.

Let

$$\mathbf{d}_{\text{dec}}(f, t) = \begin{bmatrix} d_{\text{dec},1}(f, t) \\ d_{\text{dec},2}(f, t) \\ \vdots \\ d_{\text{dec},Q}(f, t) \end{bmatrix} \quad (2.16)$$

denote a vector of mutually decorrelated signals. The decorrelated components are assumed to satisfy

$$\mathbb{E} [\mathbf{d}_{\text{dec}}(f, t) \mathbf{d}_{\text{dec}}^H(f, t)] = \mathbf{I}, \quad (2.17)$$

The goal of the covariance-domain rendering of output signals is to synthesize signals that have a covariance that closely matches a desired target covariance structure. In the current context, the target spatial characteristics are embodied in the spread diffuse coherence matrix (SDCM) that is computed from the direction of response vectors of the rendering system. To overcome this covariance matching problem, Eigenvalue decomposition (EVD) Rendering is introduced which transforms the mutually decorrelated input components into the output signals with the selected covariance structure that is desired to be a diffuse coherence.

The rendered spread signal is generated by applying a mixing matrix to the decorrelated signal vector:

$$\mathbf{d}_{\text{scm}}(f, t) = \mathbf{M}(f) \mathbf{d}_{\text{dec}}(f, t), \quad (2.18)$$

where $\mathbf{I}$ is the identity matrix. The above assumption also implies that the signal components are statistically uncorrelated and have unit covariance. The resulting decorrelated signal vector thus does not have any preprogrammed spatial coherence structure.

The covariance matrix of the rendered signal becomes

$$\mathbb{E} [\mathbf{d}_{\text{scm}}(f, t) \mathbf{d}_{\text{scm}}^H(f, t)] = \mathbf{M}(f) \mathbb{E} [\mathbf{d}_{\text{dec}}(f, t) \mathbf{d}_{\text{dec}}^H(f, t)] \mathbf{M}^H(f). \quad (2.19)$$

Substituting the identity covariance of the decorrelated components gives

$$\mathbb{E} [\mathbf{d}_{\text{scm}}(f, t) \mathbf{d}_{\text{scm}}^H(f, t)] = \mathbf{M}(f) \mathbf{M}^H(f). \quad (2.20)$$

Therefore, covariance-domain rendering becomes a covariance matching problem in which the mixing matrix must satisfy

$$\mathbf{M}(f) \mathbf{M}^H(f) = \mathbf{D}(f), \quad (2.21)$$

where $\mathbf{D}(f)$ denotes the target spread diffuse coherence matrix (SDCM).

Since $\mathbf{D}(f)$ is Hermitian and positive semi-definite, it can be decomposed using eigenvalue decomposition:

$$\mathbf{D}(f) = \mathbf{E}(f)\mathbf{\Lambda}(f)\mathbf{E}^H(f), \quad (2.22)$$

where $\mathbf{E}(f)$ is the eigenvector matrix and

$$\mathbf{\Lambda}(f) = \text{diag}(\lambda_1(f), \lambda_2(f), \dots, \lambda_Q(f)) \quad (2.23)$$

is the diagonal eigenvalue matrix. The eigenvectors describe the principal spatial directions of the covariance structure, while the eigenvalues determine the corresponding spatial energy distribution along those directions.

A direct solution to the covariance matching problem is obtained by constructing the mixing matrix as

$$\mathbf{M}(f) = \mathbf{E}(f)\mathbf{\Lambda}^{1/2}(f), \quad (2.24)$$

where $\mathbf{\Lambda}^{1/2}(f)$ denotes the square-root eigenvalue matrix:

$$\mathbf{\Lambda}^{1/2}(f) = \text{diag}\left(\sqrt{\lambda_1(f)}, \sqrt{\lambda_2(f)}, \dots, \sqrt{\lambda_Q(f)}\right). \quad (2.25)$$

The square-root eigenvalue matrix scales the decorrelated components according to the desired spatial energy distribution, while the eigenvector matrix rotates the components into the spatial subspace defined by the target covariance structure.

Substituting the EVD mixing matrix into the covariance expression yields

$$\mathbf{M}(f)\mathbf{M}^H(f) = \mathbf{E}(f)\mathbf{\Lambda}^{1/2}(f)\mathbf{\Lambda}^{1/2}(f)\mathbf{E}^H(f). \quad (2.26)$$

Since

$$\mathbf{\Lambda}^{1/2}(f)\mathbf{\Lambda}^{1/2}(f) = \mathbf{\Lambda}(f), \quad (2.27)$$

the covariance becomes

$$\mathbf{M}(f)\mathbf{M}^H(f) = \mathbf{E}(f)\mathbf{\Lambda}(f)\mathbf{E}^H(f) = \mathbf{D}(f), \quad (2.28)$$

demonstrating that the covariance of the rendered signals exactly matches the target diffuse coherence matrix.

The EVD rendering method thus carries out a full diffuse spatial rendering with a covariance structure of the inter-channel spreads that is desired between the studies. The interaural coherence of the realized binaural signals is much lower due to the decorrelation of all signal components used for the rendering. It tends to make the rendered width of the source more pronounced and blended in. Too much decorrelation on the other hand however can also impact perceptual source compactness and timbral qualities, notably for tonal or transient audio material.

2.3.3 Optimal Mixing Rendering

The rendering of EVD signals is an exact covariance matching solution, but the estimation of the resulting signals is based only on decorrelated components. This can have a profound effect on interaural coherence and can add too much diffuseness,

artefacts and perceptual colour. In particular, high deceleration can give rise to low source compactness and influence the perceptual fidelity of tonal or transient signals. The optimal mixing (OM) approach addresses this limitation by combining a prototype component with a decorrelated residual component. Instead of relying entirely on decorrelated signals, OM determines the contribution of both components through covariance-domain mixing to approximate the desired spatial covariance. The OM framework first forms a prototype binaural signal, denoted as

$$\mathbf{d}_{\text{proto}}(f, t), \quad (2.29)$$

which represents the signal rendered towards the central or prototype direction. Its covariance is given by

$$\mathbf{C}_{\text{proto}}(f) = \mathbb{E} \left[\mathbf{d}_{\text{proto}}(f, t) \mathbf{d}_{\text{proto}}^H(f, t) \right]. \quad (2.30)$$

The prototype signal provides a reference component that retains characteristics of the original source and contributes to improved spectral and spatial consistency compared with a fully decorrelated rendering.

In OM rendering, the output signal is constructed from two components: a prototype branch and a decorrelated residual branch. The rendered signal can be expressed as

$$\mathbf{d}_{\text{OM}}(f, t) = \mathbf{M}(f) \mathbf{d}_{\text{proto}}(f, t) + \mathbf{M}_r(f) \mathcal{D} [\mathbf{d}_{\text{proto}}(f, t)], \quad (2.31)$$

where $\mathbf{M}(f)$ denotes the prototype mixing matrix, $\mathbf{M}_r(f)$ denotes the residual mixing matrix, and $\mathcal{D}[\cdot]$ denotes a decorrelation operation applied to the prototype signal. The residual branch is therefore not an independent external signal, but a decorrelated version of the prototype signal. This allows the residual component to introduce additional spatial diffuseness while maintaining a relationship to the original source.

If the decorrelated residual branch is assumed to be uncorrelated with the prototype branch, the covariance of the OM-rendered signal can be written as

$$\mathbf{C}_{\text{OM}}(f) = \mathbf{M}(f) \mathbf{C}_{\text{proto}}(f) \mathbf{M}^H(f) + \mathbf{M}_r(f) \mathbf{C}_{\text{dec}}(f) \mathbf{M}_r^H(f), \quad (2.32)$$

where

$$\mathbf{C}_{\text{dec}}(f) = \mathbb{E} \left[\mathcal{D} [\mathbf{d}_{\text{proto}}(f, t)] \mathcal{D} [\mathbf{d}_{\text{proto}}(f, t)]^H \right] \quad (2.33)$$

denotes the covariance of the decorrelated prototype signal.

The OM rendering problem can then be interpreted as a constrained covariance matching problem:

$$\mathbf{C}_{\text{OM}}(f) \approx \mathbf{C}_y(f), \quad (2.34)$$

where $\mathbf{C}_y(f)$ is the target spatial covariance matrix. Unlike EVD rendering, the target covariance is not produced entirely through fully decorrelated components. Instead, the prototype branch preserves the original spatial and spectral structure as much as possible, while the residual branch compensates for the remaining covariance mismatch.

The residual covariance can be written conceptually as

$$\mathbf{C}_r(f) = \mathbf{C}_y(f) - \mathbf{M}(f) \mathbf{C}_{\text{proto}}(f) \mathbf{M}^H(f). \quad (2.35)$$

To shape the decorrelated residual component to fill in the portion of the target covariance matrix not accounted for by the prototype branch, the residual mixing matrix $\mathbf{M}_r(f)$ is used. In this interpretation the residual branch is not a substitute for the prototype branch but an addition branch that adds more spread energy to that which is spread in the prototype branch. The residual covariance should be a valid (positive semi-definite, Hermitian) covariance matrix. For practical applications, therefore, it is important to impose constraints like Hermitian enforcement, positive semi-definite projection, and so on to limit the eigenvalues to be non-negative, the covariance matrix to be symmetric, and the residual mixing to be stable.

OM rendering offers a more even trade-off between spatial diffuseness and perceptual fidelity than EVD rendering. EVD can achieve a similar target diffuse covariance structure, but relies on fully decorrelated components, which may reduce interaural coherence and introduce audible colouration. The prototype branch remains an important perceptual component in OM as the main spatial and spectral reference, since the main perceptual properties of source compactness, directional stability, and timbral consistency are mainly provided by the prototype component, while the residual branch compensates for the remaining covariance mismatch. This is advantageous for situations where excessive decorrelation results in overly diffuse spatial images or noticeable colouration.

2.4 HRTF-Based Directional Modeling

Binaural rendering systems create spatial perception by modeling the directionality of the sound that is received by the two ears. The directionality of these cues are based on the “head-related transfer function” (HRTF): it describes how an incoming sound wave from a particular spatial direction is modified by the head, torso and outer ear before reaching the listener. The HRTF thus gives the directional response characteristics needed for binaural spatial audio reproduction.

For a spatial direction represented by the unit vector $\mathbf{u}$ and a frequency band f , the binaural directional response vector may be written as

$$\mathbf{h}(f, \mathbf{u}) = \begin{bmatrix} h_L(f, \mathbf{u}) \\ h_R(f, \mathbf{u}) \end{bmatrix}, \quad (2.36)$$

where $h_L(f, \mathbf{u})$ and $h_R(f, \mathbf{u})$ denote the HRTF responses for the left and right ears, respectively. The directional response vector therefore characterises the binaural filtering associated with a single spatial direction.

In practical implementations, the directional responses are sampled over a discrete spatial grid surrounding the listener. The responses from all sampled directions may then be organised into a directional response matrix:

$$\mathbf{H}(f) = \begin{bmatrix} \mathbf{h}(f, \mathbf{u}_1) & \mathbf{h}(f, \mathbf{u}_2) & \cdots & \mathbf{h}(f, \mathbf{u}_K) \end{bmatrix}, \quad (2.37)$$

where K represents the total amount of spatial directions sampled. The columns in $\mathbf{H}(f)$ are the binaural directional responses for one direction of the spatial grid.

Covariance-domain spatial rendering is based upon the directional response matrix. The spread diffuse coherence matrix (SDCM) presented in Section 2.3.1 can be used

to numerically describe the spatial coherence structure of diffuse sound fields using the directional responses obtained from various spatial directions. An additional benefit for the discrete directional representation, is that practical covariance domain rendering can be realized with measured or modeled HRTF datasets.

The directional response matrices are calculated in the present scheme using measured head-related impulse responses (HRIRs). The first step in the processing of the HRIR signals is to convert them from the time domain to the frequency domain and the second is to arrange them in the correct spatial directions on the rendering grid. Then, these frequency-domain directional responses are combined to provide the target covariance structures for EVD and OM rendering.

The covariance-domain rendering framework relies directly on the directional response matrix, which means that the spatial characteristics of the rendered binaural signals are heavily dependent on the underlying HRTF set and spatial sampling grid. Thus, the perceived interaural coherence, spatial diffuseness and directional stability of the reproduced sound field is decided during the directional modelling stage.

2.5 Decorrelation Mechanisms

The decorrelation operation is one of the most basic operations in covariance-domain spatial audio rendering. Perceived spatial width and diffuseness of a sound source are highly dependent on the interaural coherence between the left and right ear signals in binaural reproduction. Signals that have high interaural coherence are perceived as being compact and spatially narrow, while low interaural coherence gives the wider and more diffuse perception of the reproduced sound field.

To remove the similarity between the components of the signal, one can use decorrelation in covariance-domain rendering without altering its spectral content. Thus, the decorrelation process also changes the spatial coherence structure, but does not affect the energy distribution of the underlying signals. This allows the rendering system to output a binaural version that is expanded in the spatial field, but remains perceptually the same as the original source.

For a binaural signal vector

$$\mathbf{d}(f, t) = \begin{bmatrix} d_L(f, t) \\ d_R(f, t) \end{bmatrix}, \quad (2.38)$$

The interaural coherence refers to the amount of correlation between the left and right signal components. When the signals are correlated, the resulting spatial perception is very compact, while decorrelated signals give a more diffuse spatial perception. The decorrelation process is then directly relevant to the spatial property of the binaural signals in the rendered image for covariance-domain rendering.

In EVD rendering, the signals are widened in space and are completely independent of each other. Decorrelated signals are then transformed to reflect the desired diffuse coherence structure with covariance-domain mixing matrices. With this method one can obtain the correct covariance matching and very good spatial widening, but the signals can become too diffuse if the decorrelation strength is too large.

It may also decrease perceptual source compactness, produce perceptible artefacts like colouration, unstable spatial images or lack of timbral uniformity in the case of strong decorrelation. If this is the case, these effects are more likely to be evident for tonal signals and transient audio content where too high a decorrelation level may affect the perceived spectral structure of the reproduced sound.

With OM rendering, these effects are minimized in the residual branch by only decorrelating the prototype signal. OM preserves some of the original interaural coherence of the rendered signals, but adds only the amount of decorrelated energy needed for covariance compensation. As a result, the overall spatial widening and perceptual fidelity of OM rendering tend to be higher.

Possible practical implementations of decorrelation include time varying all-pass filtering, delay based processing, phase randomisation, and other coherence reduction techniques. These methods are aimed at minimizing similarity between channels without affecting the overall perceptual properties of the signal.

The amount of decorrelation directly affects the perceptual balance between spatial diffuseness and source stability. Generally, the higher the decorrelation, the wider and more diffuse the images, the lower the decorrelation, the more compact and consistent the images. Thus, the crucial factor for covariance-domain rendering is to ensure a controlled decorrelation to obtain perceptually stable spatial widening.

3

MATLAB Implementation and System Framework

3.1 Overall Rendering Pipeline

The framework to use MATLAB for the implementation of the covariance-domain binaural source spread rendering uses multiple modules such as afSTFT domain time-frequency analysis, HRTF domain directional modeling, decorrelation processing and the covariance domain spatial mixing to render a binaural signal having perceptually widened spatial characteristics. They are processed frame by frame and covariance-domain spatial mixing is used independently in each time-frequency region, producing tight spatial control and allowing a flexible adaptation based on the input signals.

The pipeline starts with the input audio being divided into overlapping frames, and the audio in each of these frames is converted to the time-frequency domain using an alias-free short-time Fourier transform (afSTFT) framework. This representation breaks the signal up into its frequency-dependent components and allows for good covariance-domain spatial processing as well as good reconstruction with weighted, overlap-add synthesis. This framework incorporates the mechanisms of directionality, partial decorrelation, and covariance-domain mixing to generate binaural outputs that reproduce spatial spread and localization cues in the auditory scene.

Following time-frequency analysis, binaural prototype signals are generated through HRTF-based directional modeling. The directional responses obtained from an HRTF dataset encapsulate the spatial filtering characteristics associated with desired source directions, and thus serve as foundational cues for spatial rendering. These prototype signals establish the perceptual basis for subsequent covariance-domain processing, encoding directional localization information that is critical for accurate externalization and azimuthal perception in binaural reproduction [11], [12].

A parallel decorrelated signal branch is subsequently derived from the prototype signals using decorrelation processing. The decorrelated components are essential for introducing spatial diffuseness and reducing coherence between channels, which supports perceptual widening of the auditory scene [5]. In the context of eigenvalue decomposition (EVD) rendering, the output signals are synthesized exclusively from the decorrelated signal components, prioritizing diffuse spatial attributes over compact directional focus. In contrast, optimal mixing (OM) rendering combines an optimised prototype contribution with a decorrelated residual branch, providing a

compromise between covariance matching and perceptual source preservation. This dual-branch architecture facilitates a continuum between focused and diffuse spatial impressions, enabling flexible control over spatial attributes [5].

During the covariance-domain mixing stage, frequency dependent mixing matrices are used to create binaural signals whose spatial covariance is matched to the desired diffuse coherence structure. Target covariance matrices are computed from the directional response model developed by the HRTF framework, and specify the desired spatial spread features of the reproduced sound field. This method takes advantage of the fact that covariance matrices can serve as a spatial descriptor, and interaural coherence and cross-channel covariance of the spatialized signals are matched to perceptual criteria of spatial envelopment [4]. The application of covariance targets further allows for perceptually motivated control of spatial attributes such as diffuseness, apparent source width and externalization.

Covariance-domain mixing is followed by afSTFT inverse synthesis and overlap-add reconstruction into the time domain. The weighted overlap-add method guarantees the continuity of time between neighboring frames, to achieve smooth frame transitions and prevent artifacts in reconstruction. This reconstruction method is crucial for processing a stable binaural representation from subsequent processing blocks and for ensuring perceptual continuity in non-stationary stimuli like speech and music [13].

The overall rendering framework consequently comprises of a directional modeling, a decorrelation processing, a covariance domain spatial mixing and a time-frequency reconstruction into a single binaural rendering system. Modular structure allows for control over spatial widening without compromising perceptual stability or binaural consistency of the spatial dimension. Moreover, the covariance-domain representation facilitates perceptually relevant spatial manipulation by coherence control that is frequency dependent, supporting the rendered auditory scene to be both directionally accurate and spatially enveloping. These features are crucial for an immersive binaural reproduction system and virtual auditory environments.

The covariance-domain framework can also be used to support perceptual control of apparent source width and interaural coherence in addition to the spatial widening. Previous studies have demonstrated that interaural correlation (IC) manipulation for each frequency band can affect perceived envelopment and spatial impression in binaural listening situations. The framework could be extended by adaptive covariance targets and listener-dependent HRTFs datasets, enabling dynamic rendering in head tracked and/or individual listening situations [8].

3.2 afSTFT Processing Framework

The covariance-domain rendering framework[4], [14] operates in the time-frequency domain using a WOLA-based afSTFT processing structure. The afSTFT framework provides frame-based frequency-domain representations suitable for covariance-domain spatial processing while maintaining stable reconstruction through weighted overlap-add synthesis.

In the present implementation, the input signals are processed independently for each frame and channel. Let $x_q[n]$ denote the time-domain input signal for channel

q . During frame-based processing, the incoming samples are first appended to the analysis buffer while preserving the overlap from the previous frame. The buffered frame is then multiplied by the analysis window:

$$x_q^{\text{win}}(n, t) = x_q^{\text{buf}}(n, t) w_a[n], \quad (3.1)$$

where $x_q^{\text{buf}}(n, t)$ denotes the buffered signal frame at time frame t and $w_a[n]$ is the analysis window.

The windowed frame is subsequently transformed into the frequency domain using the discrete Fourier transform:

$$X_{f,q}(t) = \text{FFT} \left\{ x_q^{\text{win}}(n, t) \right\}, \quad (3.2)$$

where $\text{FFT}\{\cdot\}$ denotes the Fourier transform operation. The resulting frequency-domain coefficients are then used for covariance-domain rendering and binaural spatial processing.

In the MATLAB implementation, the FFT analysis is performed independently for each channel after updating the overlap-add analysis buffer. Only the non-redundant frequency bands are retained for subsequent covariance-domain processing. The resulting frequency-domain coefficients are organised into frame-wise time-frequency representations for each processing block.

To improve low-frequency resolution, the afSTFT framework additionally incorporates hybrid-band processing for the lower frequency bins. The hybrid-band decomposition follows the afSTFT formulation introduced in Chapter 2 and generates additional low-frequency subbands for improved spatial processing stability while maintaining efficient reconstruction performance.

After covariance-domain rendering, the processed frequency-domain coefficients are transformed back into the time domain using inverse afSTFT synthesis. Let $\hat{X}_{f,q}(t)$ denote the processed frequency-domain coefficients after covariance-domain mixing. The reconstructed time-domain frame is obtained through inverse Fourier transformation:

$$\hat{x}_q(n, t) = \text{IFFT} \left\{ \hat{X}_{f,q}(t) \right\}, \quad (3.3)$$

where $\text{IFFT}\{\cdot\}$ denotes the inverse Fourier transform operation.

The reconstructed frame is then multiplied by the synthesis window and combined with neighbouring frames using weighted overlap-add reconstruction:

$$y_q[n] = \sum_t \hat{x}_q(n, t) w_s[n - tR], \quad (3.4)$$

where $w_s[n]$ denotes the synthesis window and R represents the hop size between adjacent frames.

The afSTFT processing framework therefore provides a stable time-frequency representation suitable for covariance-domain spatial rendering. By combining frame-based spectral analysis, hybrid-band decomposition, and overlap-add reconstruction, the framework enables frequency-dependent binaural covariance processing while maintaining perceptually stable signal reconstruction.

3.3 HRTF Processing and Directional Grid Construction

The covariance-domain rendering framework requires directional binaural responses in order to construct the target spatial covariance structures used for spatial widening. In the present implementation, the directional responses are obtained from measured head-related impulse responses (HRIRs), which are subsequently transformed into the frequency domain for covariance-domain processing [14], [15].

The HRIR dataset contains binaural impulse responses measured for multiple spatial directions surrounding the listener. Each directional measurement consists of a left-ear and right-ear impulse response together with the corresponding spatial coordinates. During initialisation, the HRIR data are loaded into the rendering framework and validated to ensure consistency between the directional grid, sampling rate, and binaural channel configuration. The implementation supports SOFA-compatible HRIR datasets for binaural directional modelling.

Since the rendering framework operates at a fixed processing sampling rate, the HRIR dataset is resampled when necessary to maintain consistency between the input audio signals and the directional response data. This ensures that the afSTFT processing framework and the HRTF responses operate within the same sampling-rate domain.

The spatial directions associated with the HRIR measurements are represented using azimuth and elevation coordinates. For covariance-domain processing, the directional grid is additionally converted into Cartesian unit vectors:

$$\mathbf{u} = \begin{bmatrix} x \\ y \\ z \end{bmatrix}, \quad (3.5)$$

where $\mathbf{u}$ denotes the direction-of-arrival unit vector associated with a particular spatial direction on the rendering grid.

The HRIR responses are subsequently transformed into the frequency domain using Fourier analysis in order to construct the directional response vectors required for covariance-domain rendering. For frequency band f and spatial direction $\mathbf{u}$, the binaural directional response vector is represented as

$$\mathbf{h}(f, \mathbf{u}) = \begin{bmatrix} h_L(f, \mathbf{u}) \\ h_R(f, \mathbf{u}) \end{bmatrix}, \quad (3.6)$$

where $h_L(f, \mathbf{u})$ and $h_R(f, \mathbf{u})$ denote the left-ear and right-ear frequency-domain HRTF responses, respectively.

The directional response vectors from all sampled directions are then organised into the directional response matrix

$$\mathbf{H}(f) = [\mathbf{h}(f, \mathbf{u}_1), \mathbf{h}(f, \mathbf{u}_2), \dots, \mathbf{h}(f, \mathbf{u}_K)], \quad (3.7)$$

where K denotes the number of sampled spatial directions on the rendering grid. The directional response matrix provides the binaural spatial response structure required for covariance-domain rendering and diffuse coherence estimation.

To improve computational efficiency during rendering, the covariance-domain implementation additionally precomputes directional outer-product matrices of the form

$$\mathbf{h}(f, \mathbf{u}_k)\mathbf{h}^H(f, \mathbf{u}_k), \quad (3.8)$$

for each spatial direction $\mathbf{u}_k$ on the rendering grid. These precomputed matrices are subsequently reused during covariance construction and diffuse coherence estimation, reducing repeated matrix calculations during frame-wise rendering.

In the MATLAB implementation, the HRTF processing framework therefore consists of HRIR loading, directional-grid construction, frequency-domain directional response generation, directional response matrix construction, and covariance pre-computation. These stages provide the spatial response structures required for both EVD and OM covariance-domain rendering.

3.4 EVD Rendering Implementation

The EVD rendering framework is implemented using covariance-domain mixing based on the target diffuse coherence matrices constructed from the HRTF directional response framework. The implementation follows the covariance-domain formulation introduced in Chapter 2 and performs rendering independently for each frequency band and time frame.

For each frequency band f , the target diffuse coherence matrix is represented as

$$\mathbf{D}(f). \quad (3.9)$$

The covariance-domain mixing matrices are obtained through eigenvalue decomposition:

$$\mathbf{D}(f) = \mathbf{E}(f)\mathbf{\Lambda}(f)\mathbf{E}^H(f), \quad (3.10)$$

where $\mathbf{E}(f)$ contains the eigenvectors of the covariance matrix and $\mathbf{\Lambda}(f)$ is the diagonal matrix of eigenvalues.

The covariance-domain mixing matrix is then constructed as

$$\mathbf{M}(f) = \mathbf{E}(f)\mathbf{\Lambda}^{1/2}(f). \quad (3.11)$$

In the MATLAB implementation, the eigenvalue decomposition is performed independently for each frequency band using the target covariance matrices generated from the directional response framework. The resulting mixing matrices are subsequently applied to decorrelated binaural signal components during frame-wise rendering.

The decorrelated binaural signal vector is denoted as

$$\mathbf{d}_{\text{dec}}(f, t). \quad (3.12)$$

The rendered binaural signals are then obtained through covariance-domain mixing:

$$\mathbf{d}_{\text{scm}}(f, t) = \mathbf{M}(f)\mathbf{d}_{\text{dec}}(f, t). \quad (3.13)$$

The covariance-domain mixing matrix is constructed such that the covariance of the rendered binaural signals approximately matches the target diffuse coherence structure described in Chapter 2. Consequently, the EVD rendering framework provides strong spatial widening through fully decorrelated binaural signal generation.

Since the covariance matrices must remain Hermitian and positive semi-definite, numerical safeguards are applied during covariance estimation in order to maintain stable decomposition behaviour. The covariance-domain rendering is then performed independently for each time-frequency frame within the afSTFT processing structure.

The EVD rendering framework represents a direct covariance matching approach in the present implementation. Because the rendered signals are generated entirely from decorrelated components, the method generally produces strong spatial diffuseness and reduced interaural coherence. However, excessive decorrelation may also reduce perceptual source compactness and increase audible coloration artefacts for tonal or transient audio material [5].

3.5 Optimal Mixing Rendering

Optimal mixing (OM) rendering is realised based on the covariance-domain optimisation method proposed in the reference method. OM has a prototype component with high fidelity and a residual decorrelated component to correct for the difference in covariance, unlike the EVD rendering method, where decorrelated signals are used only to decorrelate the covariance.

The prototype signal is first generated using the central spreading direction. In the time-frequency domain, the prototype signal is expressed as

$$\mathbf{d}_{proto}(f, t) = \mathbf{h}_{cen}(f)s(f, t), \quad (3.14)$$

where $\mathbf{h}_{cen}(f)$ represents the HRTF corresponding to the central source direction and $s(f, t)$ denotes the input source signal.

The covariance matrix of the prototype signal is estimated as

$$\mathbf{C}_{proto}(f) = \mathbb{E}[\mathbf{d}_{proto}(f, t)\mathbf{d}_{proto}^H(f, t)]. \quad (3.15)$$

In the OM framework, the prototype component is first transformed through an optimised mixing matrix $\mathbf{M}(f)$:

$$\mathbf{d}_{OM}(f, t) = \mathbf{M}(f)\mathbf{d}_{proto}(f, t) + \mathbf{M}_r(f)\mathcal{D}[\mathbf{d}_{proto}(f, t)], \quad (3.16)$$

where $\mathcal{D}[\cdot]$ denotes the decorrelation operation and $\mathbf{M}_r(f)$ represents the residual mixing matrix.

The primary mixing matrix is obtained by solving a covariance-domain optimisation problem. The objective is to minimise the signal distortion while satisfying the covariance constraint:

$$\mathbf{M}\mathbf{C}_{proto}\mathbf{M}^H + \mathbf{M}_r\mathbf{C}_{dec}\mathbf{M}_r^H = \mathbf{C}_y. \quad (3.17)$$

Following the optimal mixing formulation, the primary mixing matrix is calculated by decomposing the target covariance and prototype covariance matrices and solving the corresponding optimisation problem. The resulting matrix can be expressed as

$$\mathbf{M} = \mathbf{K}_y \mathbf{P} \mathbf{K}_{proto}^{-1}, \quad (3.18)$$

where $\mathbf{K}_y$ and $\mathbf{K}_{proto}$ are obtained from covariance decompositions, and $\mathbf{P}$ is determined through singular value decomposition of the optimisation term.

After applying the primary mixing matrix, the remaining covariance mismatch is calculated as

$$\mathbf{C}_r = \mathbf{C}_y - \mathbf{M} \mathbf{C}_{proto} \mathbf{M}^H. \quad (3.19)$$

The residual covariance is then used to construct the residual mixing matrix $\mathbf{M}_r$, which introduces only the additional decorrelated energy required to reach the target covariance. This avoids applying full decorrelation to the entire signal and therefore reduces possible coherence loss and perceptual artefacts compared with EVD rendering.

During implementation, covariance matrices are regularised and constrained to remain Hermitian and positive semi-definite before eigenvalue decomposition. The obtained mixing matrices are interpolated between consecutive processing frames to achieve smooth time-frequency transitions.

Compared with EVD rendering, OM provides a better compromise between covariance matching and signal preservation by combining an optimised prototype component with a controlled amount of decorrelated residual energy.

4

Results and Discussion

4.1 Evaluation Methodology

4.1.1 Evaluation Signals and Rendering Conditions

The evaluation is conducted using several categories of audio stimuli, including white noise, male speech, female speech, music-related material, seagull sounds, and ocean wave recordings. These stimuli are selected in order to evaluate both spatial rendering behaviour and perceptual signal fidelity under different spectral and temporal characteristics.

The rendering evaluation is performed for source spreading angles of 0° , 30° , 60° , 90° , 120° , 150° , and 180° . For each spreading condition, binaural signals are generated using the baseline (BL), unconstrained EVD, and OM rendering methods.

Reference signals correspond to the prototype binaural rendering with no source spreading. The rendered binaural outputs are then compared against the reference signals using perceptual metrics and signal distortion measures.

4.1.2 Binaural Covariance Representation

For a given frequency band f , the rendered binaural signal vector is defined as

$$\mathbf{d}(f) = \begin{bmatrix} d_l(f) \\ d_r(f) \end{bmatrix}, \quad (4.1)$$

where $d_l(f)$ and $d_r(f)$ denote the left and right binaural channels, respectively.

The covariance matrix of the binaural signal is defined as

$$\mathbf{C}_d(f) = \mathbb{E} [\mathbf{d}(f)\mathbf{d}^H(f)] = \begin{bmatrix} c_{d,ll}(f) & c_{d,lr}(f) \\ c_{d,rl}(f) & c_{d,rr}(f) \end{bmatrix}, \quad (4.2)$$

where $c_{d,ll}(f)$ and $c_{d,rr}(f)$ represent the auto-spectral powers of the left and right channels, while $c_{d,lr}(f)$ and $c_{d,rl}(f)$ represent the cross-spectral terms between the binaural channels.

The covariance representation provides a compact description of the binaural spatial characteristics and forms the basis for the perceptual metrics used in the evaluation.

4.1.3 Perceptual Metrics

Three binaural perceptual parameters are considered in the evaluation: colouration, interaural level difference (ILD), and interaural coherence (IC).

The colouration metric is defined as

$$\text{Colouration}(f) = 10 \log_{10} [c_{d,u}(f) + c_{d,rr}(f)]. \quad (4.3)$$

The interaural level difference is defined as

$$\text{ILD}(f) = 10 \log_{10} \frac{c_{d,u}(f)}{c_{d,rr}(f)}. \quad (4.4)$$

The interaural coherence is defined as

$$\text{IC}(f) = \frac{\text{real}[c_{d,lr}(f)]}{\sqrt{c_{d,u}(f)c_{d,rr}(f)}}. \quad (4.5)$$

The IC metric describes the similarity between the left and right binaural channels. Higher IC values generally correspond to a more compact spatial image, while lower IC values indicate increased spatial diffuseness. The ILD metric represents the binaural level difference between the ears and is an important cue for horizontal localisation. The colouration metric reflects the overall spectral energy distribution of the rendered binaural signal.

4.1.4 RMSE-based Metric Evaluation

The perceptual metrics are evaluated by comparing the rendered binaural signals against the reference binaural signals. For each frequency band, the difference between the rendered metric and the corresponding reference metric is computed. The overall error is then quantified using the root-mean-square error (RMSE):

$$\text{RMSE} = \sqrt{\frac{1}{K} \sum_{k=1}^K (x_k - x_{\text{ref},k})^2}, \quad (4.6)$$

where x_k denotes the evaluated perceptual metric at frequency band k , $x_{\text{ref},k}$ denotes the corresponding reference metric, and K represents the total number of frequency bands.

The RMSE evaluation quantifies how closely the rendered binaural cues match the reference binaural cues for different spreading conditions and rendering methods.

4.1.5 Signal Distortion Error

In addition to frequency-domain perceptual metrics, waveform fidelity is evaluated using the signal distortion error (SDE):

$$\text{SDE} = \sqrt{\frac{\sum_n \|\mathbf{d}(n) - \mathbf{d}_{\text{ref}}(n)\|^2}{\sum_n \|\mathbf{d}_{\text{ref}}(n)\|^2}}, \quad (4.7)$$

where $\mathbf{d}(n)$ denotes the rendered binaural signal and $\mathbf{d}_{\text{ref}}(n)$ denotes the reference binaural signal.

The SDE metric measures the overall waveform deviation between the rendered and reference signals in the time domain. Lower SDE values correspond to smaller signal distortion and better waveform fidelity.

4.2 Objective Evaluation Results

4.2.1 Frequency-Dependent Binaural Metric Analysis

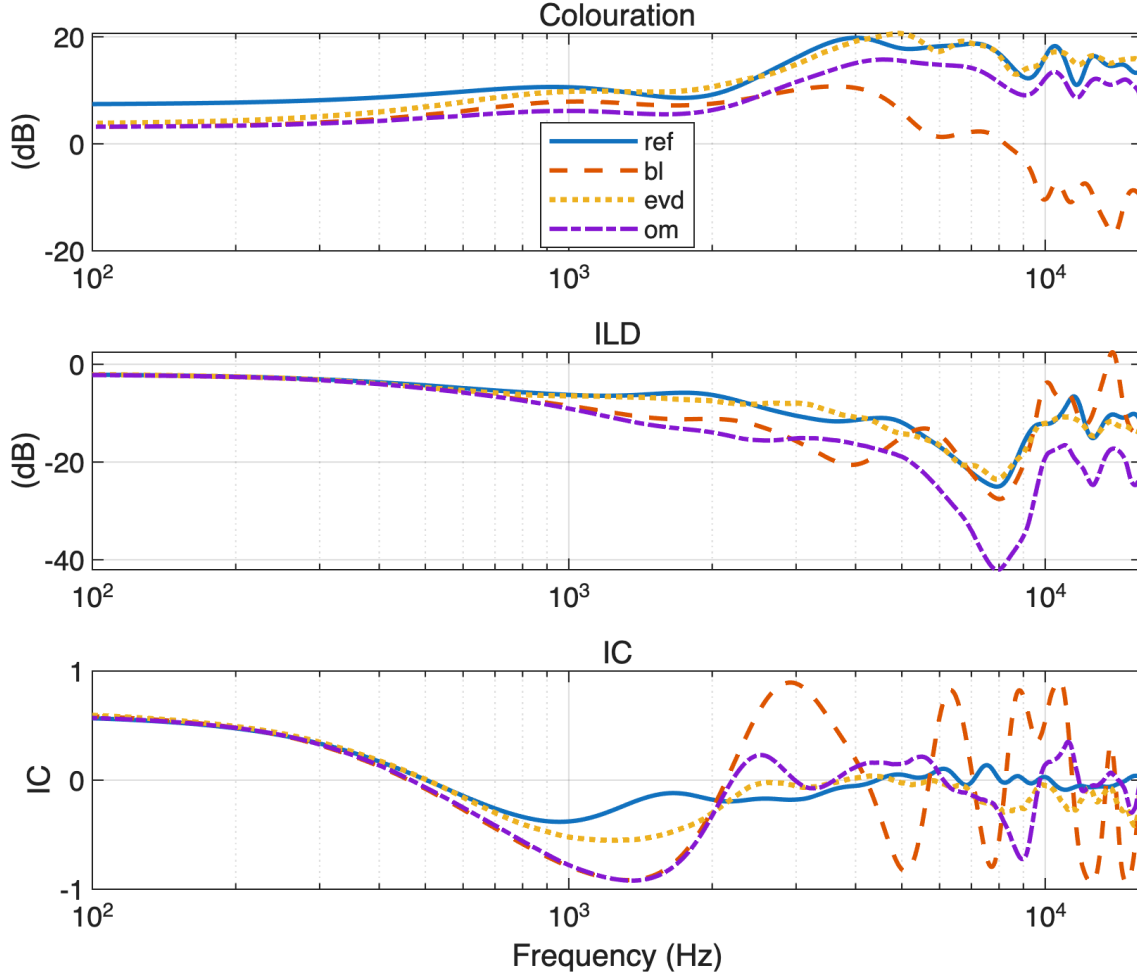


Figure 4.1: Frequency-dependent binaural metrics for the reference, BL, EVD and OM renderings using the generated white noise stimulus. The plots show Colouration, ILD and IC across the frequency range for a source direction of 45° and a source spread of 60° .

The white noise signal used for the evaluation is a 10 second monophonic signal generated in MATLAB, which was used consistently for all rendering methods. All white noises were generated with a fixed random seed to assure that they were the same for all experiments. Before rendering it binaurally, the signal was DC corrected, RMS normalised and scaled to prevent clipping. The binaural metrics for the rendered outputs were then evaluated using frequency-dependent binaural metrics, such as Colouration, interaural level difference (ILD), and interaural coherence (IC).

Figure 4.1 presents the frequency-dependent evaluation results of the reference signal and the three rendering methods: BL, EVD, and OM. The comparison is made for

the frequency range between 100 Hz and 16 kHz for spectral deviation, binaural level differences, and interaural coherence.

The results of the Colouration indicate that the response of EVD is closest to the reference response in the majority of frequencies. This means that the covariance matching process can be used to replicate the desired spatial covariance properties without altering significantly the spectral properties. The trend of the reference signal is followed by OM, with larger variations in the high-frequency range. For BL, the response strongly different from the reference can be seen for frequencies higher than approximately 10 kHz, which is referred to as the spectral mismatch. This behaviour is expected since BL is based on coherent rendering and does not explicitly optimise the output covariance towards the target spread distribution.

The ILD results show different characteristics compared with Colouration. The relationship between EVD and the reference signal is relatively good throughout the frequency range. Despite the optimisation of the prototype and residual mixing matrices, the covariance matching, however, does not result in smaller deviations in the ILD for OM, especially in the middle and high frequency regions. Matching the spatial covariance matrix does not guarantee the exact reproduction of all binaural cues, which means that the matched spatial covariance matrix may not correspond to the same attributes of the desired spatial sound. BL follows a similar ILD trend at lower frequencies, while larger fluctuations are observed at higher frequencies, because of the difficulty in coherent rendering for reproducing the target spread characteristics.

The results obtained with the IC are used to illustrate the effect of the different decorrelation strategies in the following. The EVD is closest to the reference coherence behaviour as predicted by its fully decorrelated covariance matching formulation. In OM, the final coherence behaviour is determined by the interaction between the prototype and residual components. As a result, OM exhibits an intermediate IC behaviour between BL and EVD, reflecting a compromise between coherent signal preservation and covariance-based coherence reduction. The greatest offset of the reference IC response is measured for BL, for which the coherent rendering approach is unable to reproduce the reduced coherence due to the spatially extended source.

4.2.2 Spread-Dependent Metric Analysis

Figure 4.2 presents the average RMSE values and standard deviations of Colouration, ILD and IC over all 836 directions for different source spread angles. The results demonstrate how the rendering performance changes as the target spread increases from 0° to 180° . In general, increasing source spread leads to larger deviations from the reference signal for all rendering methods, although the behaviour differs significantly between BL, EVD and OM.

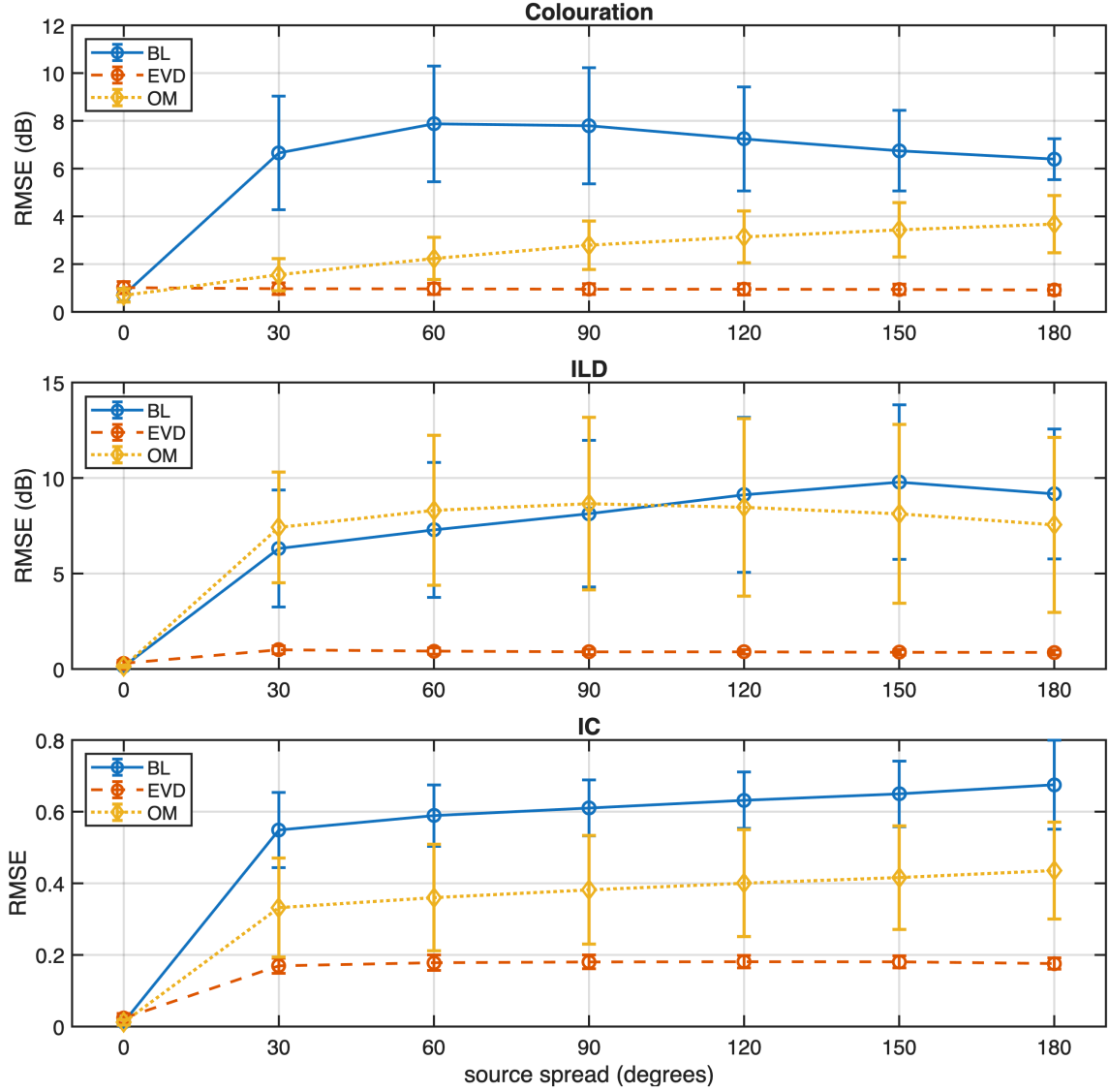


Figure 4.2: Average RMSE values and standard deviations of Colouration, ILD and IC for BL, EVD and OM renderings over all 836 directions for different source spread angles.

Colouration results indicate significant differences among three rendering methods for various spreads. The error is relatively stable for increased spreading of the source and EVD has the lowest error for all in hand spread angles. This behaviour is related to the direct covariance matching formulation of EVD where the mixing matrix is derived from the target covariance matrix. Also, Colouration errors are relatively low in OM, but the error slowly rises for larger spreads. This behavior is a combination of the optimized prototype and the remaining components: it is a spectral response that is different from that of the reference rendering. BL, in contrast, shows a steep growth of Colouration error starting from 30° and a higher error level for higher spreads. This means coherent rendering is not as suitable for wide-spread sources to reproduce its spectral characteristics.

The characteristics observed in the ILD results differ from those of Colouration. EVD achieves the lowest error for nearly all spread angles, indicating accurate re-

production of the binaural level relationship in the present implementation. This behaviour is mainly related to the covariance matching process employed in EVD, which directly optimises the output spatial covariance towards the target distribution. In contrast, BL relies on coherent rendering and cannot accurately reproduce the binaural level characteristics of an extended source, resulting in increasing ILD errors as the source spread increases.

The ILD deviations of OM are higher than BL for moderate spread angles between 30° and 90° , while becoming lower than BL for larger spreads. Although the prototype and residual mixing matrices are optimised to approximate the target covariance structure, matching the spatial covariance does not necessarily guarantee identical reproduction of individual binaural cues such as ILD. The observed behaviour reflects the influence of the prototype and residual components in OM, where the covariance optimisation modifies the binaural level distribution in order to reproduce the characteristics of spatially extended sources.

The IC result shows the various spatial characteristics of the three approaches. EVD appears to yield the lowest IC error, agreeing with its full decorrelation of the covariance matching strategy and indicating it is better able to replicate the coherence reduction of the reference signal. The greatest IC error is associated with BL, where interaural correlation is higher and spatial diffuseness lower, due to coherent rendering. The contribution from the optimised prototype signal and the decorrelated residual signal are both included in the final output signal of OM. This enables extra spatial spreading with part of the features of the original signal preserved.

The error in rendering from the various source directions is shown by the standard deviation bars. BL generally shows a greater degree of variation in the directions with the exception of Colouration and ILD which show more variation in BL, while EVD shows more consistent behaviour across the directions evaluated. OM provides an intermediate change of direction (IOD) based on the metric considered. The major reason for these variations is due to the different HRTF response for different directions during the binaural rendering process.

4.2.3 Spatial Deviation Error Analysis

Table 4.2 presents the SDE results for BL, EVD and OM using different audio stimuli. The SDE metric evaluates the deviation of the rendered spread signals relative to the compact reference source located at 90° with 0° spread. Lower SDE values therefore indicate better preservation of the original signal structure and spatial characteristics.

Table 4.1: SDE results for different stimuli and rendering methods.

Stimuli	BL	EVD	OM
Birds	0.4323	1.0706	0.4221
Drums	0.0759	4.8707	1.1068
Male speech	0.1964	1.7516	0.3849
Female speech	0.2631	1.2690	0.1828
Waves	0.3407	1.3756	0.7403
Average	0.2621	2.1548	0.5674

The SDE results clearly demonstrate the differences between the rendering methods and the influence of various covariance domain processing methods. The highest SDE values for all stimuli are produced by EVD with an average value of 2.1548. This means that the covariance matching process when fully decorrelated yields the greatest modification of the original compact source signal. This strong decorrelation which is done while performing EVD is good for spatial diffuseness and coherence reduction but it alters more of the temporal and spatial structure of signal.

The average value of SDE is smallest for BL, as in this case no decorrelation processing takes place during rendering. BL ensures higher similarity of the original signal, while preserving the coherent structure of it, with only limited spatial spreading in comparison with the covariance-based renderings. The SDE for OM is intermediate average at 0.5674. By combining the optimised prototype with the residual mixing components, OM provides extra spatial spreading capabilities and reduces the signal deviation relative to fully decorrelated EVD approach.

The stimulus-dependent behaviour also underscores the importance of the stimulus characteristics on the rendering performance. A significant rise in SDE is seen for the drum stimulus, especially for EVD which has a value of 4.8707. This behaviour is associated with the transient properties of drum signals, which are characterized by steep temporal variations, and so are more sensitive to decorrelation processing. Due to the fact that EVD is only based on decorrelated components, the transient structure may be altered to a greater extent, which leads to a higher difference from the compact original source signal. However, increasing the SDE value for drums can also be seen in OM (1.1068), yet the deviation is significantly lower compared to EVD, thanks to the optimised prototype component.

For all other stimuli, the SDE values are in the expected order with larger deviations for EVD compared to OM and BL. The SDE for speech signals is larger for OM than BL, especially for speech from male speakers, representing measurable differences between OM and BL associated with the added spatial processing provided by OM. It can be seen that for waves, BL and OM are closer together than for transient signals, meaning that smoother and more diffuse signals are less affected by the extra processing using covariance.

The SDE results are also indicative of the spatial spreading versus signal preservation dilemma. EVD has the greatest decorrelation and coherence modification, and the greatest deviation from the original source structure. BL has the best signal fidelity with little spatial widening. An intermediate solution is provided by OM, which fuses prototype and residual parts in a covariance domain optimisation, thereby

getting further spatial spreading without the signal distortion of fully decorrelated rendering.

4.3 Comparison with the Original Results

4.3.1 Frequency-Dependent Metric Comparison

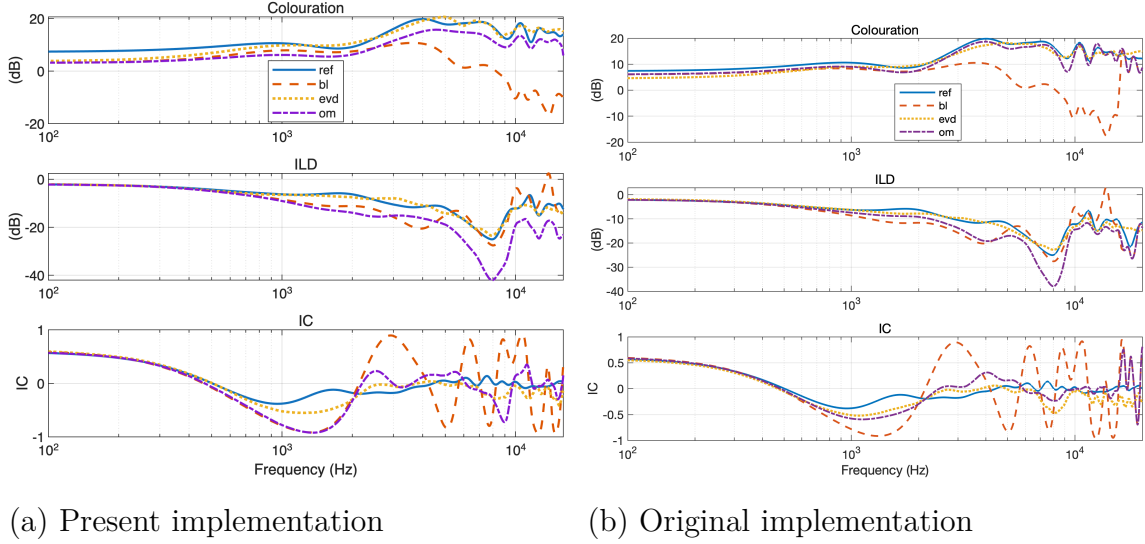


Figure 4.3: Comparison of frequency-dependent Colouration, ILD and IC responses between the present implementation and the original implementation for a source located at 45° with 60° spread.

Figure 4.1 presents the frequency-dependent Colouration, ILD, and IC responses obtained from the present MATLAB implementation, while Figure 4.3 shows the corresponding results from the original implementation. The two implementations are compared in terms of their frequency-dependent behaviour to evaluate the consistency of the reproduced covariance-domain source spread rendering framework. The BL results show similar overall behaviour between the two implementations. Comparable trends are seen in the ILD and IC responses in both implementations, and the coherent rendering characteristics are preserved. The largest difference is observed in the high-frequency Colouration response, where the present implementation shows a bigger deviation from the reference signal. Because of the coherent addition of the directional HRTFs in BL rendering, slight differences in the signal processing chain can produce perceptible differences in the spectrum, especially at high frequencies.

The results obtained in EVD are seen to have good agreement between the two implementations. Similar frequency dependent behaviour is seen with the Colouration, ILD and IC responses, and the covariance matching formulation and the eigenvalue based mixing process has been consistently reproduced. The close agreement between the three binaural metrics indicates a good reproduction of the EVD rendering framework in the present implementation. The small differences left are primarily

due to numerical discrepancies arising from the estimation of the covariance and implementation of the signal processing.

Results from the OM rendering exhibit similar behaviour between the two implementations as well. The prototype and the residual components are both computed using the optimal mixing framework by covariance domain optimisation for both implementations. The differences are primarily seen in the Colouration response at certain frequencies and in the IC response below about 2 kHz. Colouration deviations are high-frequency; small differences in the decorrelation and covariance processing can lead to a difference in the spectral behaviour of the residual component. The difference in low-frequency IC is related to the coherence characteristics introduced by the residual branch because OM is a combination of the optimised prototype contribution and a decorrelated residual branch. However, the general trends with Colouration, ILD and IC stay the same as with the original implementation.

The comparison demonstrates that the present MATLAB implementation successfully reproduces the main characteristics of the original covariance-domain source spread rendering framework. The remaining differences are mainly associated with implementation-specific aspects, including the decorrelation process, covariance estimation, and numerical signal processing, rather than fundamental differences in the rendering algorithms.

4.3.2 Global RMSE Comparison Across Source Spreads

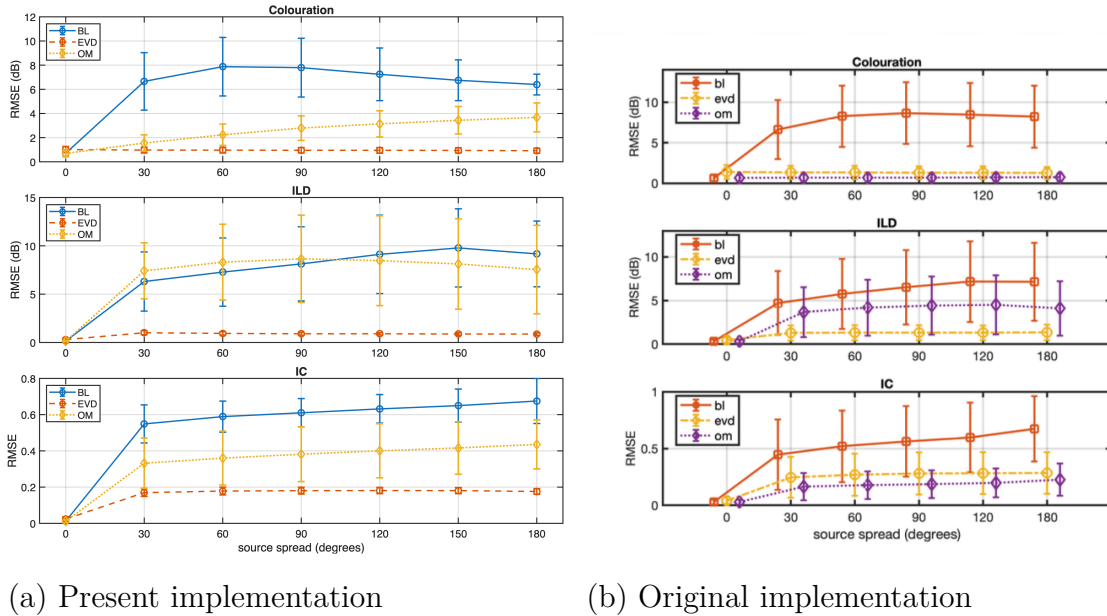


Figure 4.4: Comparison of global RMSE results for Colouration, ILD and IC across different source spreads between the present implementation and the original implementation.

Figure 4.2 and Figure 4.4 compare the spread-dependent RMSE results of Colouration, ILD and IC obtained from the present implementation and the original imple-

mentation, respectively. Since the original implementation is provided as a real-time plugin and cannot be evaluated under exactly the same conditions as the present MATLAB implementation, the original results are used for qualitative comparison of the main trends.

For Colouration, both implementations show that BL produces the largest error, and the error increases with increasing source spread. For the original implementation, EVD and OM these errors are kept relatively low over the spread range. This is the behaviour that is replicated in the present implementation, where the Colouration error of EVD is lower while that of OM is slightly higher. The difference between EVD and OM is due to the different way of signal generation. EVD generates the output from the covariance-matched decorrelated components directly, while OM is able to combine the prototype and the residual components via covariance-domain optimisation. The OM also includes a residual component, causing a little spectral modification, and thus higher Colouration error than EVD. In both cases, the much more prominent Colouration error of BL is due to the coherent addition of the directional HRTFs which sensitivity to the spectral changes becomes more prominent as the source spread widens.

The same holds true for the ILD results: there is a similar tendency between both implementation, but there are some differences in the relative magnitude. The implementation in the original mode is the one that obtains the lowest error in the ILD, BL and OM present larger deviations. The present implementation is able to reproduce the low ILD error of EVD, for all spread angles. In the case of BL, the ILD error grows steadily as the source spread is increased in both cases, and this is related to the limited capabilities of coherent rendering to replicate the binaural level characteristics of an extended source. The behaviour of OM is slightly different in the present implementation: the ILD error is higher than BL for moderate spread angles between 30° and 90° , while becoming lower than BL for larger spreads. This behaviour is related to the interaction between the optimised prototype and residual branches. in which the covariance-domain optimisation is changing the binaural level distribution to approximate a wider spread.

The best agreement between the two implementations is for the IC results. In both, EVD exhibits the least IC error, BL the greatest error and OM in between. The regularity of this behavior is an assurance that the most important coherence properties of the covariance-domain rendering framework have been reproduced. By using fully decorrelated components the lower IC error of the EVD is achieved by the fact that higher coherence reduction is allowed, while BL maintains the higher IC because of its coherent rendering structure. The OM signals have an intermediate coherence behaviour as they are composed of prototype signal information and a decorrelated residual part.

The standard deviation bars indicate the differences in the directionality of the rendering methods. The present implementation shows greater deviations than the original implementation, especially in the ILD metric, for OM. This indicates stronger direction-dependent variation in the current OM implementation, which may result from the interaction between the optimised prototype and residual components. In contrast, EVD exhibits relatively stable behavior over varying source directions, while BL is more sensitive to the directional variation of HRTFs, because of coherent

summation.

4.3.3 Spatial Deviation Error Comparison

Table 4.2: SDE results for different stimuli and rendering methods.

Stimuli	BL	EVD	OM
Birds	0.4323	1.0706	0.4221
Drums	0.0759	4.8707	1.1068
Male speech	0.1964	1.7516	0.3849
Female speech	0.2631	1.2690	0.1828
Waves	0.3407	1.3756	0.7403
Average	0.2621	2.1548	0.5674

Table 4.3: SDE results of the original implementation.

Stimuli	BL	EVD	OM
Birds	0.4676	1.2767	0.4617
Drums	0.0538	1.2222	0.0554
Male speech	0.1722	1.3533	0.1794
Female speech	0.2560	1.4244	0.2602
Waves	0.3165	1.4094	0.3294
Average	0.2532	1.3372	0.2572

Table 4.2 and Table 4.3 compare the SDE results of the present implementation and the original implementation. The SDE metric measures the deviation of the rendered spread signal from the compact reference source. Therefore, lower SDE values indicate that more of the original signal structure is preserved.

Comparing the SDE shows that in both implementations, EVD has the highest error compared to the compact reference source. The average SDE obtained from EVD is 2.1548 for the present implementation compared to 1.3372 for the original implementation. The structure of EVD is fully decorrelated, meaning that the output signal is produced only with covariance matched decorrelated components. This approach helps to achieve high spatial coherence reduction and bring more diffusedness to the original signal, but it also adds greater temporal and spatial changes to the original signal.

The main difference between the two implementations is observed in the OM results. The average SDE for OM in the original implementation is 0.2572, which is very near to the BL value of 0.2532. In comparison, the present implementation has an average SDE of 0.5674, which is still quite low, but higher than BL and lower than EVD. This difference is not due to a different OM formulation since both implementations use the optimal mixing framework. Rather, it is mostly due to differences in the covariance estimation and residual decorrelation processing that is used in its implementation. Since the OM output is made up of prototype and

residual parts, the characteristics of the residual branch can also affect the overall signal deviation while preserving the target covariance.

The stimulus-dependent results further illustrate the influence of signal characteristics on the rendering performance. The largest deviation is observed for the drum stimulus, particularly for EVD, where the SDE reaches 4.8707 in the present implementation compared with 1.2222 in the original implementation. This behaviour is related to the transient characteristics of drum signals, which contain rapid temporal variations and are therefore more sensitive to decorrelation and phase modifications. Since EVD relies entirely on decorrelated components, transient signal structures can be modified more significantly, resulting in larger deviations from the compact reference source. A similar tendency can also be observed for OM, where the SDE of drums reaches 1.1068, although the deviation remains considerably lower than EVD due to the contribution of the prototype component.

An additional analysis was done excluding the drum signal to further investigate the effect of the drum stimulus on the average SDE results. The average SDE values for present implementation are 0.3081, 1.3667 and 0.4325 for BL, EVD and OM, respectively, without the drum stimulus. These results are much closer to the original, especially for EVD which has a much smaller difference between the two (1.3667 vs 1.3660). This indicates that the large difference in the average EVD result is mainly caused by the stronger sensitivity of transient signals to the applied decorrelation process, rather than a fundamental difference in the rendering framework. However, the smaller deviation of the OM after drums removal also reveals that the residual part of the OM has more impact on signals with quick temporal changes.

The differences between the two implementations, for the other stimuli, is stimulus dependent. For the bird stimulus, there is a good agreement between the two implementations, with comparable SDE values for OM. In the present implementation, the OM SDE is slightly lower for female speech and higher for male speech than in the original implementation, which suggest that the effect of the residual component is dependent on the temporal aspects of the input signal for speech signals. The wave stimulus, on the other hand, indicates a higher difference for OM (SDE of 0.7403 in the present implementation as opposed to the original implementation of 0.3294). This means that for even smoother signals, the deviation of the signal can vary as a result of difference in covariance estimation and residual decorrelation processing.

5

Conclusion and Future Work

A MATLAB implementation of covariance-domain source spread rendering for binaural audio reproduction was presented in this thesis. The implementation included baseline coherent rendering (BL), eigenvalue decomposition (EVD) and optimal mixing (OM) in accordance with the spatial covariance matching framework proposed by McCormack et al. The perceptually widened binaural signals were obtained by combining HRTF-based directional modelling, afSTFT time-frequency processing, covariance estimation, decorrelation and covariance-domain mixing.

Colouration, interaural level difference (ILD), interaural coherence (IC) and spatial deviation error (SDE) were used to assess the implemented rendering techniques. The basic properties of covariance-domain source spread rendering were reproduced when compared to the original implementation. The strongest coherence modification and spatial diffuseness was obtained by generating the output from covariance-matched decorrelated components using EVD, whereas using OM provided a balance between covariance matching and preservation of the original signal structure.

The evaluation results also revealed different trade-offs introduced by the rendering strategies. EVD was able to reproduce the target coherence behaviour most closely but with greater deviations from the compact source signal especially for transient programme material. BL maintained the original coherent signal structure, offering only limited control over the spatial spreading. OM achieved an intermediate behaviour by combining prototype and residual components through covariance-domain optimisation, allowing additional spatial widening while preserving more of the original signal characteristics.

Generally very similar frequency and spread dependent behaviour compared with the original implementation. Other quantitative differences were mostly for specific stimuli and/or metrics, such as transient signals in the SDE evaluation. Analysis without the drum stimulus revealed that these discrepancies were highly correlated to the sensitivity of transient signals to decorrelation processing but not to basic differences in the covariance domain rendering framework as such.

There are a number of limitations in the current implementation. The assessment was largely made via objective measurements and informal listening tests; formal listening tests were not conducted. Also, the decorrelation/covariance has been done independent of the original real-time plugin, and thus can have numerical and implementation differences.

Future research may include further perceptual stability and reducing the difference between offline MATLAB rendering and real time implementation further. Extensions may be adaptive decorrelation strategies, transient aware covariance domain processing, perceptually motivated compensation strategies and wider range subjec-

tive listening evaluations. Moreover, if the implementation could be optimised for real-time processing it would be possible to use covariance-domain source spread rendering in interactive spatial audio production settings.

Bibliography

- [1] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization*, Revised Edition. Cambridge, MA, USA: MIT Press, 1997.
- [2] D. R. Begault, *3-D Sound for Virtual Reality and Multimedia*. Academic Press, 1994.
- [3] V. Pulkki, “Virtual sound source positioning using vector base amplitude panning”, *Journal of the Audio Engineering Society*, vol. 45, no. 6, pp. 456–466, 1997.
- [4] V. Pulkki, “Spatial sound reproduction with directional audio coding”, *Journal of the Audio Engineering Society*, vol. 55, no. 6, pp. 503–516, 2007.
- [5] L. McCormack, A. Politis, and V. Pulkki, “Rendering of source spread for arbitrary playback setups based on spatial covariance matching”, in *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2021, pp. 371–375. DOI: 10.1109/WASPAA52581.2021.9632724.
- [6] G. Potard and I. Burnett, “Decorrelation techniques for the rendering of apparent sound source width in 3d audio displays”, in *Proceedings of the 7th International Conference on Digital Audio Effects (DAFx)*, 2004, pp. 1–6.
- [7] C. Jenny and C. Reuter, “Usability of individualized head-related transfer functions in virtual reality: Empirical study with perceptual attributes in sagittal plane sound localization”, *JMIR Serious Games*, vol. 8, no. 3, e17576, 2020. DOI: 10.2196/17576.
- [8] E. M. Wenzel, M. Arruda, D. J. Kistler, and F. L. Wightman, “Localization using nonindividualized head-related transfer functions”, *The Journal of the Acoustical Society of America*, vol. 94, no. 1, pp. 111–123, 1993. DOI: 10.1121/1.407089.
- [9] S. Fernandez, L. McCormack, A. Politis, and V. Pulkki, “Enhancing binaural rendering of head-worn microphone arrays using spatial covariance matching”, *The Journal of the Acoustical Society of America*, vol. 151, no. 4, pp. 2618–2634, 2022. DOI: 10.1121/10.0010109.
- [10] R. Nordahl and N. C. Nilsson, “The sound of being there: Presence and interactive audio in immersive virtual reality”, in *The Oxford Handbook of Interactive Audio*, Oxford University Press, 2014. DOI: 10.1093/oxfordhb/9780199797226.013.013.

- [11] F. L. Wightman and D. J. Kistler, “Headphone simulation of free-field listening. i: Stimulus synthesis”, *The Journal of the Acoustical Society of America*, vol. 85, no. 2, pp. 858–867, 1989. DOI: 10.1121/1.397557.
- [12] A. Kulkarni, S. K. Isabelle, and H. S. Colburn, “Sensitivity of human subjects to head-related transfer-function phase spectra”, *The Journal of the Acoustical Society of America*, vol. 105, no. 5, pp. 2821–2840, 1999.
- [13] R. E. Crochiere, “A weighted overlap-add method of short-time fourier analysis/synthesis”, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 1, pp. 99–102, 1980. DOI: 10.1109/TASSP.1980.1163353.
- [14] V. Pulkki, S. Delikaris-Manias, and A. Politis, Eds., *Parametric Time-Frequency Domain Spatial Audio*. Wiley, 2017. DOI: 10.1002/9781119252634.
- [15] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, “The cipic hrtf database”, in *Proceedings of the 2001 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2001, pp. 99–102. DOI: 10.1109/ASPAA.2001.969552.

A

Core MATLAB Implementation Snippets

A.1 EVD Covariance-Domain Mixing

The EVD rendering implementation performs covariance-domain mixing through eigenvalue decomposition of the target covariance matrix. Negative eigenvalues are clipped in order to maintain numerical stability and physically valid covariance behaviour.

```
[V, D] = eig(C);  
  
lambda = real(diag(D));  
lambda(lambda < 0) = 0;  
  
M = V * diag(sqrt(lambda));
```

A.2 OM Covariance-Domain Mixing

The OM rendering implementation obtains both the prototype mixing matrix and the residual mixing matrix through covariance-domain optimisation. First, the prototype mixing matrix is calculated to match the contribution of the prototype covariance matrix to the target covariance. The remaining covariance difference is then represented by the residual covariance matrix, from which the residual mixing matrix is obtained.

```
% Prototype covariance optimisation  
[M, Cr] = spr_formulate_M_and_Cr_cmplx( ...  
    Cproto, Cy, Qmix_cmplx, ...  
    useEnergyFLAG, reg);  
  
% Residual covariance mixing  
[Mr, ~] = spr_formulate_M_and_Cr_real( ...  
    CprotoDiag, Cr, Qmix, ...  
    useEnergyFLAG, reg);
```


B

Rendering and Evaluation Parameters

Table B.1: Rendering parameters used in the implementation.

Parameter	Value
Sampling frequency	48 kHz
Frame size	512
Hop size	128
Hybrid bands	133
Number of ears Q	2
HRTF dataset	Provided by the original implementation framework
Spread angles	$0^\circ - 180^\circ$
Evaluation directions	836
Test signals	White noise, speech, birds, drums, waves

Table B.2: Evaluation parameters used for objective metrics.

Parameter	Value
STFT window length	4096
FFT size	4096
Hop size	1024
Window type	Hann
Metrics	Colouration, ILD, IC, SDE

DEPARTMENT OF SOME SUBJECT OR TECHNOLOGY
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden
www.chalmers.se



CHALMERS
UNIVERSITY OF TECHNOLOGY